Project Thesaurus 2020

Sylvie Reusch, University of Burgundy (France)

Keywords: Thesaurus, linguistics, concepts, ontologies

Abstract

Linguistic thesauri are analyzed and strategies for the future development of such thesauri are discussed. Then linguistic thesauri are linked to other kinds of thesauri used in computer sciences, knowledge management , and further areas. Experiences from all these areas contribute to the design of future thesauri. A project to develop the concepts for a future thesaurus has been start - Project 2020.

1. Introduction

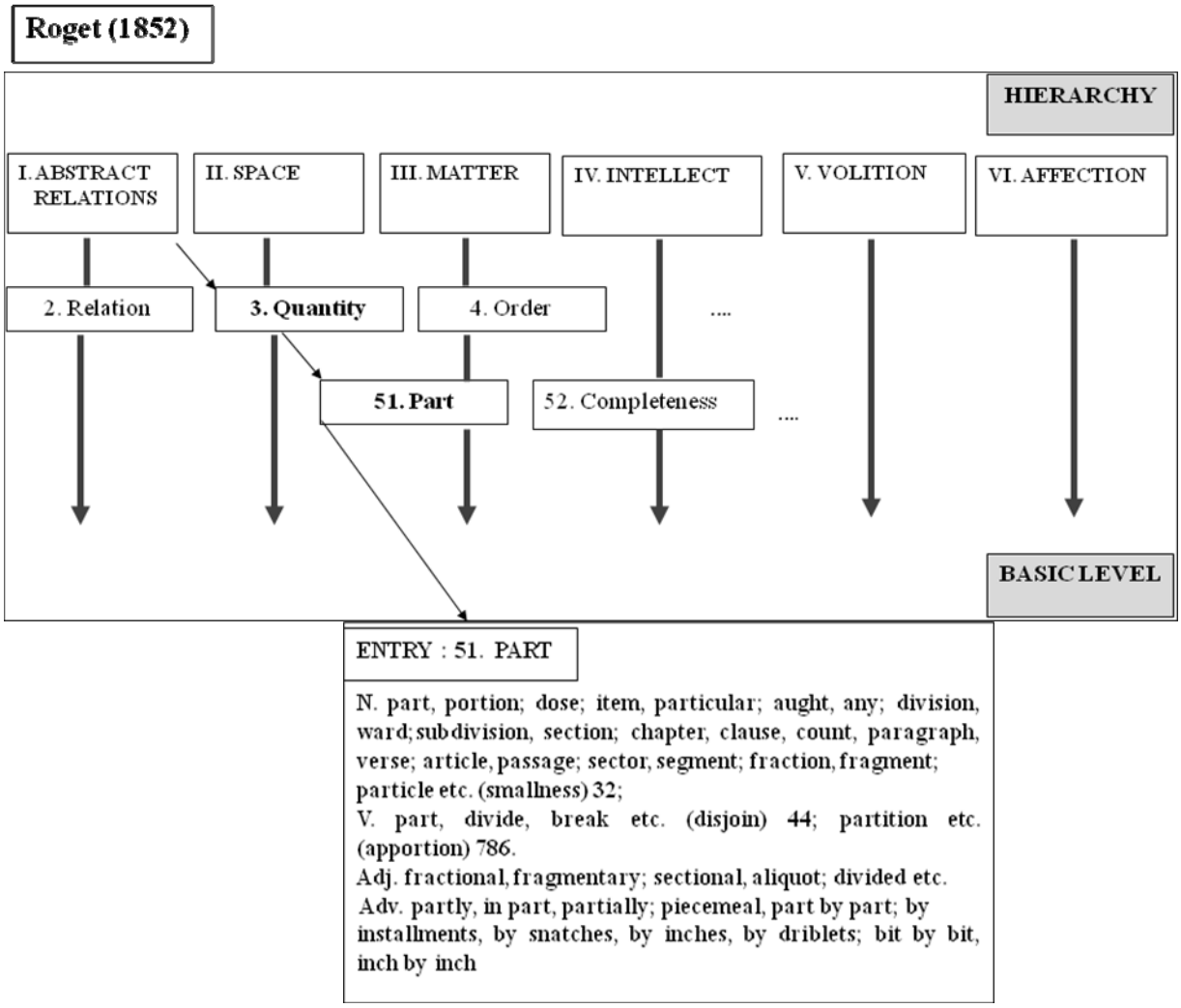
A linguistic thesaurus consists of a classificatory system that introduces concepts (*preferred terms*) in a hierarchy, with several relations used to link synonyms and otherwise related terms, and additional parts with terms in alphabetical order or further explanations.

Today, thesauri are used to learn languages, to classify documents, to support creativity and for many more reasons. Thesauri are available all over the world. Peter Mark Roget's *"Thesaurus of English words and phrases"*, the prototype for modern linguistic thesauri, first published in 1852 [1], has a new edition every year since more than 150 years. But today we have a crisis on thesauri. Some new publications of Roget's Thesaurus suffer from conceptual rearrangements, that destroy the powerful link between words and concepts. For the future it is important to strengthen the classical approach and develop it across the languages. And in addition to linguistical concepts of computer science and knowledge management can stimulate the development of thesauri.

2. Linguistic Thesauri

A linguistic thesaurus generally consists of two main parts: The *Conceptual Hierarchy*, where words are arranged according to their *signification,* and an *Alphabetical Index*, where words are arranged in alphabetic order. The conceptual hierarchy of Roget's Thesaurus is composed of six primary classes: *Abstract Relations, Space, Matter, Intellect, Volition, Affections* - as shown in figure 1. Each category is composed of divisions and sections, where concepts are arranged in a hierarchy.

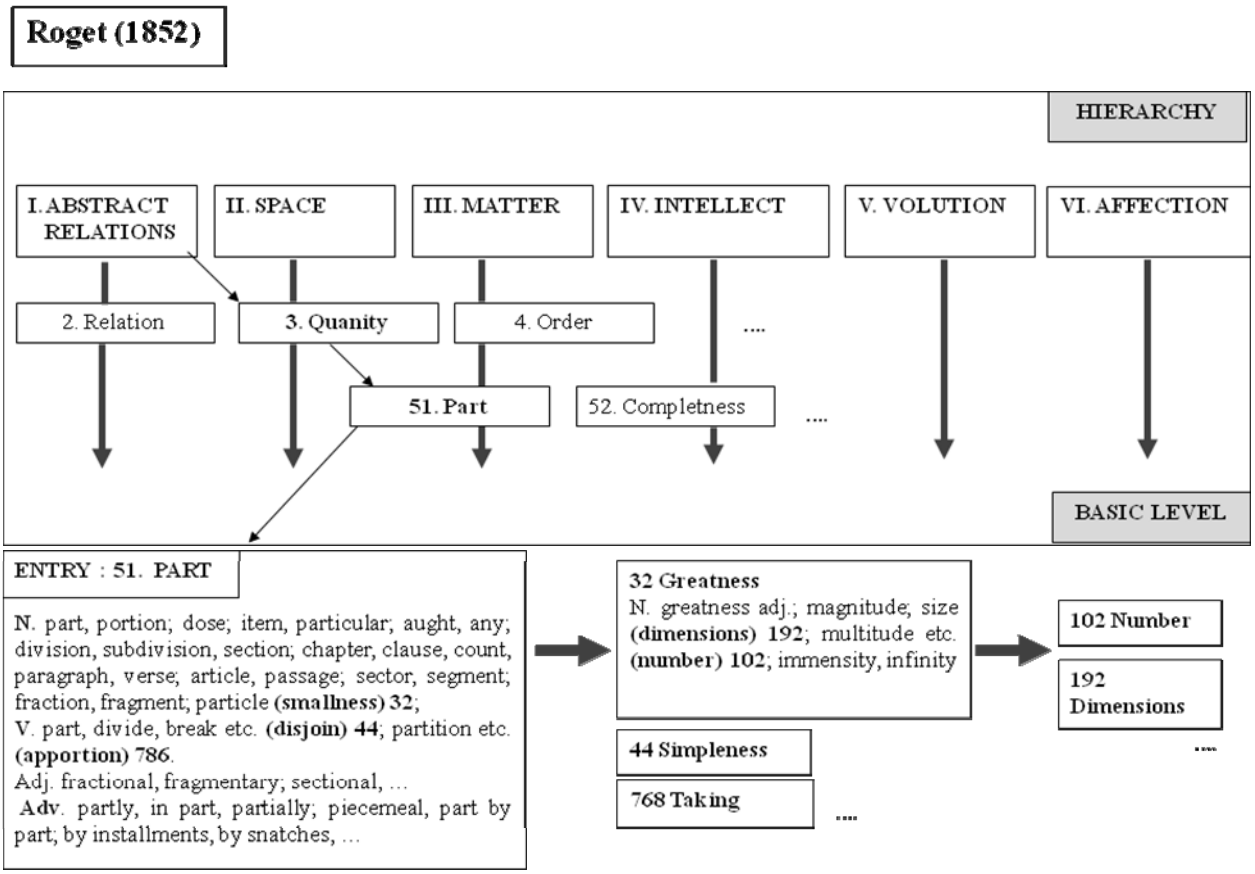
Figure 1: The Structure of Roget's Thesaurus



At the lowest level of the hierarchy in Roget´s thesaurus there are about 1000 concepts - concept 51 *Part* and concept 52 *Completeness* shown in figure 1 are two of these. For each such concept at the lowest level of the conceptual hierarchy there is an entry consisting of a set of nouns, adjectives, verbs and other terms. Figure 1 shows that entry for concept 51 *Part*. The words listed here represent near synonyms to a preferred term. Such an entry can therefore be seen as a spectrum of a concept.

The entries include references to entries of other concepts as shown in figure 2. From the entry of concept 51 (*Part*) to entry 32 (*Greatness/Smallness*), further to entry 102 (*Number*), etc. .

Figure 2: Cross references in Roget's Thesaurus



Peter Mark Roget believed that his classificatory system, based on the organization of concepts, could be aligned to any kind of language: *"The principles of its construction are universally applicable to all languages, wheter living or dead. On the same plan of classification there might be formed a French, a German, a Latin, or a Greek Thesaurus, possessing, in their repective spheres, the sames advantages as those of the English model."* [Roget, 1852, p. xxii]. In fact, soon after publication, Roget´s Thesaurus was recognized as a milestone for dictionaries and was adopted to various European languages [Hüllen, 2009, p.60]. Thesauri based on Roget´s Thesaurus were developed in Germany, Spain and France, for example. In France, "*Le dictionnaire idéologique*" (1858) of Théodore Robertson, published with the authorization of Mark Roget, is the first application of the Roget´s approach to French language [Hüllen, 2009, p.84]. Robertson wished to stay as close as possible to the admired original as shown in table 1. In the introduction he confesses: *"I have been engaged for years in collecting and classifying material* [*for a topical dictionary*]*. By inspecting the publication of Mr Roget I recognized that the collection* [*of words*] *which must serve as the basis for my work has already been completed: that it is more perfect than the*

one which I had worked on with so much pains, and that it was far superior as regards classification." [translated by Hüllen, 2009, pp. 77/78]. Therefore, the French thesaurus must be seen as a literal translation of the admired English original.

Table 1 Roget and Robertson. Comparing the conceptual hierarchy

Roget (1852)			**Robertson (1859)**		
I. Abstract relations	1. Existence	2. Relation	**I. Rapports abstraits**	1. Existence	2. Relation
	3. Quantitiy	4. Order		3. Quantité	4. Ordre
	5. Number	6. Time		5. Nombre	6. Temps
	7. Change	8. Causation		7. Changement	8. Causalité
II. Space	1. Generally	3. Form	**II. Espace**	1. En général	2. Forme
	2. Dimensions	4. Motion		2. Dimensions	4. Mouvement
III. Matter	1. Generally	2. Inorganic	**III. Matière**	1. En général	2. Inorganique
	3. Organic			3. Organique	
IV. Intellect	1. Formation of ideas	2. communication of ideas	**IV. Intelligence**	1. Formation des idées	2. Communication des idées
V. Volition	1. Individual	2. Intersocial	**V. Volonté**	1. Individuelle	2. Mutuelle
VI. Affections	1. Generally	2. Personal	**VI. Affections**	1. En général	2. Personnelles
	3. Sympathetic	4. Moral		3. Sympathiques	4. Morales
	5. Religious			5. Religieuses	

Robertson did not only translate the conceptual hierarchy of Roget´s Thesaurus, but even when writing the text entries at the basic level he "followed each entry article word for word when writing the English translation" [Hüllen, 2009, p.86] - as shown in table 1.

Table 2 Roget and Robertson. Comparing the entry "*Existence*"

Roget (1852)	**Robertson (1859)**
I ABSTRACT RELATIONS **1 Existence** **N. Existence, being, entity, ens[Lat], esse[Lat], reality, actuality; positiveness fact, matter of fact, truth 494; Science of existence, ontology.** **V. exist, be; have being; subsist, live, breathe, stand,be the case; occur; have place, find oneself, remain, stay. …**	**I RELATIONS ABSTRAITES** **1** **Existence** **1. EXISTENCE, être, entité, *ens*, moi. Coexistence (120).** **Réalité, actualité, l'absolu, fait, état de consistance. Vérité (494).** **Science de l'être, Ontologie.** **Existence dans l'espace (186).** ***V.* Être, exister, subsister, respirer, vivre, être sur terre, avoir lieu, se trouver, se rencontrer, se conserver, rouler, y avoir, il y a. …**

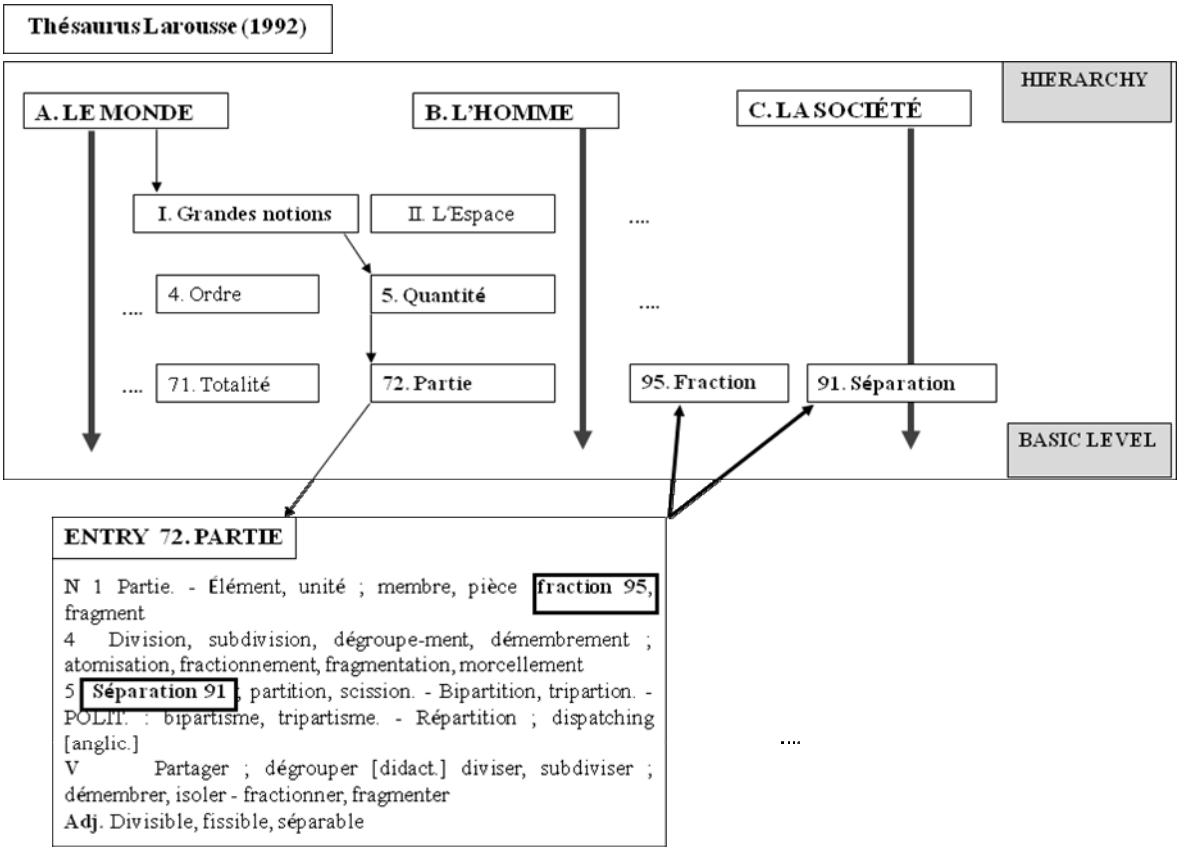
Other European lexicographers were more eager in developing their own thesaurus approach. A famous example for Germany is Daniel Sander's *Deutscher Sprachschatz (1877)* [10].

Roget had a strong influence on lexicography in the 19th and at the beginning of the 20th century. *"Although it is itself the outcome of a long and complex European tradition, the particular form in which it appeared 1852 was so convincing that after about one hundred years almost every European language had its own thesaurus."* [Hüllen, 2009, p.89]

But in the middle of the 20th century French lexicographers ignored Théodore Robertson´s Dictionnaire idéologique. In the French speaking world thesauri were replaced by dictionaries of synonyms that do not use a hierarchy of concepts, analogy or other term relations [4].

In 1992 Larousse decided to create a *new* French thesaurus. Aware of the unbroken succes of Roget in the English World the publisher wanted to create a *Thésaurus Larousse*. The Thésaurus Larousse does not follow Roget´s Thesaurus. Thésaurus Larousse includes a hierarchy of concepts that is totally different from Roget´s Thesaurus as shown in figure 3. Here we have three main categories: *World, Human, Society* [3].

Figure 3: The functionality of the Thésaurus Larousse



All three categories contain a number of subcategories often composed of categories which are derived from Roget's hierarchy. For example, the category B *Human* consists of the sub-categories *Generality, Body, Mind, Volition and Action*. Roget's higher-ordered category *Volition* is integrated in the hierarchy on a lower level (in the category B-IV). The classification system of the Thésaurus Larousse is composed of an higher number of sub-categories than the *Thesaurus of English Words*. The starting point of the classification is the philosophical Trinity *World, Human, Society*. The reduction of concepts at the top level of the hierarchy is balanced by the introduction of a new categories on a second and third level of hierarchy as shown in Table 3.

Table 3: The classification system of the Thésaurus Larousse

A. LE MONDE			
B. L'HOMME	**I. Généralités**	1. L'individu	2. Les âges de la vie
	II. Le corps	1. parties du corps 3. Preception 5. Santé et maladie	2. fonctions corporelles 4. États du corps 6. médécine
	III. L'esprit	1. Intelligence 3. Vie spirituelle	2. Affectivité
	IV. La volonté et l'action	1. Volonté	2. Action
C. LA SOCIÉTÉ	**I. Les relations sociales**	1. Rapports entre personnes	2. Hiérarchies 3. Conflits
	II. La vie collective	1. Organisation sociale 3. Droit	2. Morale
	III. La communication	1. Généralités 3. Supports et vecteurs de la communication	2. Signes et sens 4. Arts
	IV. Les métiers et les activités	1. Emploi 3. Agriculture et pêche 5. Échanges économiques	2. Industrie 4. Transports

There are problems in the hierarchy of concepts in Thésaurus Larousse regarding the labeling of categories. The following figure shows that concepts C.II.1.c. and entry 678 have the same label. When concepts and entries do not have a unique label it is very hard to discuss that - it´s a source for misunderstanding.

Figure 4: Problems of labeling categories

THESAURUS LAROUSSE (1992)

C. LA SOCIÉTÉ

II. VIE COLLECTIVE

1. Organisation sociale

c. Famille

678 Famille
679 Père
680 Mère

Other problems appear when we look at the hierarchy of concepts.

Figure 5: Structural problems in the conceptual hierarchy

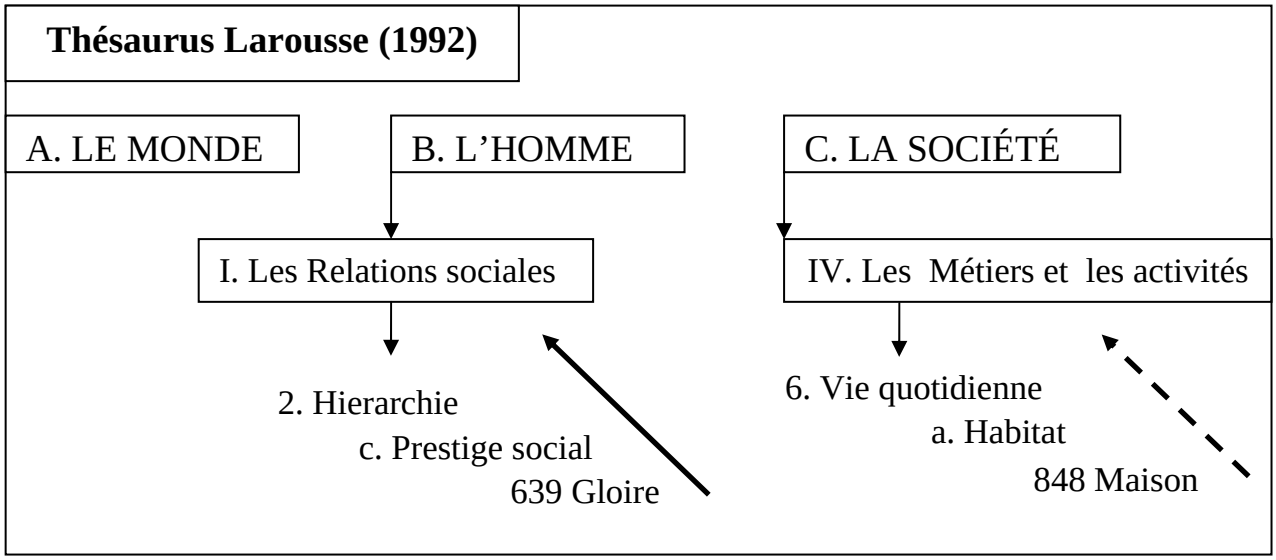


Figure 5 shows 2 paths in the conceptual hierarchy of Thésaurus Larousse.

The first case shows a good example for classification. The entry *Glory* appears in the sub-category *Social Prestige* that is included in *Hierarchy* and is ranged in the main category *Social Relations*. We can see a logic progression: *Glory* is a kind of *Social prestige* that is determined by an existing *Hierarchy*. A *Hierarchy* is a special kind of *Social relation* that is based on the existence of concept B on *Human*.

The second case shows us a bad example for classification. The entry *House* is located in the category *Everyday life* itself included in a category labeled *Profession and Hobbies.* We can see a logic break: *House* is a part of *Everyday life,* but is not a part of the category IV *Professions and hobbies.* The logic break is even more visible when creating a mind map for the conceptual field *Everyday life.* The structure of a mind map would underline that general Objects (like a house) and general Human activities (like a hobby) create separate branches, not necessarily depending on each other. More exemples for classification problems can be found [17]. For example, the entry *Architecture* appears in a category labeled *Communication*. The entry *Mathematics* is arranged under ! a category labeled *Number* - on the same level as entries like *One, Two, Three*. In this case, a number is no longer considered as a mathematical object used to count and measure, but as super-category to Mathematics.

The concept hierarchy of Thésaurus Larousse is much weaker than the concept hierarchy of Roget´s Thesaurus. Thésaurus Larousse also failed from an economical point of view. No new edition followed the first edition of 1992.

So lessons learned from more than 150 years of linguistic thesauri are:

- Thesauri like Roget´s Thesaurus are well established and recognized.
- The conceptual hierarchy of Roget´s Thesaurus is useful for learning and discussing concepts even today.
- Roget´s approach should be further developed to meet future requirements regard the linguistic perspectives.

In the next sections we discuss further perspectives of thesauri, that will also shape the future of thesauri.

3. Further Kinds of Thesauri and International Standards

Besides linguistic thesauri there are thesauri used in computer science and for knowledge representation. Thesauri in these areas follow the same approach of concept hierarchies and establish similar links among concepts. In contrast to linguistic thesauri these thesauri are mainly used to describe documents and to support information retrieval. The IEEE thesaurus [13] or the Inspec Thesaurus [12] are such thesauri. As usual for all core concepts used in computer science and related areas there is an International Standards for Thesauri (ISO 25964-1).

The ISO Standard [11] redefines the objective and the nature for modern linguistic thesauri in the following way: *"The tradition aim of a thesaurus is to guide the indexer and the searcher to choose the same term for the same concept. In order to achieve this, a thesaurus should first list all the* ***concepts that might be useful for retrieval purposes*** *in a given domain. The concepts are represented by terms, and for each concept, one of the possible representations is selected as the preferred term. Secondly, a thesaurus should* ***present the preferred term in such a way that people will easily identify the one(s) they need****. This is achieved by establishing relationships between terms - and/or between concepts - and using the relationships to present the terms in a structured display."* [ISO25964-1, p.13].

4. Combined Approaches for future Thesauri

Future thesauri can be based upon the concepts of linguistic and non-linguistic thesauri as discussed so far. Beyond that, descriptions of concepts can be improved by mathematical approaches - by concept lattices for example [14]. Algorithms are available for the manipulation of concept lattices [14].

Thesauri can be implemented based on topic maps [15 - 18]. Topic maps can support many kinds of links among concepts and terms. Topic maps are appropriate to support linguistical concepts and information retrieval [17,18].

5. Conclusions

A review of classical linguistical approaches, the recognition of international standards, and further mathematical and IT-based approach can be used to developed a thesaurus that meets future requirements - Thesaurus 2020. The master thesis of the author and further linked projects will support the initiative.

6. References

[1] Roget, Mark. *Thesaurus of English words and Phrases*. 150 th Anniversary Edition edited by George Davidson, London 2002.

[2] Robertson, Théodore. *Dictionnaire idéologique. Recueil des mots, des phrases, des idiotismes et des proverbes de la langue française classés selon l'ordre des idées,* Paris London 1859.

[3] Péchoin, Daniel. *Thésaurus Larousse. Des idées aux mots et des mots aux idées*, Paris 1992.

[4] Quemada, Bernard. *Les Dictionnaires du Francais moderne 1539 - 1863. Etude sur leur histoire, leurs types et leurs méthodes*, Paris 1967.

[5] Steger, Hugo (et al.). *Dictionaries. An International Encyclopedia of Lexicography*. New York 1989

[6] Péchoin, Daniel. *Thésaurus documentaire, thésaurus idéologique et dictionnaire analogique.* In: Cotten, J.-P.;Hufschmitt B. Repérer, formaliser, traduire les concepts philosophiques, Colloque de Besancon 23-24 juin 1999, Paris 2001.

[7] Monneret, Philippe. *Essais de linguistique analogique*, Dijon 2004

[8] Bianucci, Gabriele [u.a.] *"L'usage de relations sémantiques dans l'élaboration de thesauri: L'expérience du PTP (Petit Thesaurus Politique),"* in: Cahiers de lexicologie LXI:2 , 1992, 59-84.

[9] Hüllen, Werner. *Networks and knowledge in Roget's Thesaurus. From ancient to medieval.* Oxford 2009.

[10] Sanders, Daniel: Deutscher Sprachschatz, Hamburg 1873-1877, reprint Tübingen 1985.

[11] BSI Group headquarters, *ISO 25964-1 Information and documentation - Thesauri and interoperability with other vocablaries - Part 1 : Thesauri for information retrevial*, London 2009.

[12] IET Inspec Thesaurus 2010 - www.theiet.org

[13] IEEE Thesaurus 2009 - http://www.ieee.org/documents/2009Taxonomy_v101.pdf

[14] Kovács, László : Concept Lattice Structure with Attribute Lattice, http://www.bmf.hu/conferences/HUCI2003/kovacs.pdf.

[15] www.ontopia.net

[16] Meschede, Michael: Ontologien und Topic Maps zur Unterstützung des Konferenzmanagement, Diplomarbeit Dortmund 2011.

[17] Reusch, Sylvie: Des idées aux mots, de mots aux idées. Étude de la structure et du fonctionnement d'un thésaurus francais. Dijon 2011 (planned).

[18] Reusch, Sylvie: Project Thesaurus 2020 - Linguistic and Ontological Aspects, Conference on Intelligent Data Acquisition and Advanced Computing Systems – IEEE-IDAACS 2011 - Prague. 2011 (submitted).